FEBS Letters 586 (2012) 962–966

FEBS Letters

journal homepage: www.FEBSLetters.org

Hypothesis

Protein folding by 'levels of separation': A hypothesis

Lesley H. Greene [a,*], Terri M. Grant [b]

[a] Department of Chemistry and Biochemistry, Old Dominion University, Norfolk, VA 23529, USA
[b] Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529, USA

ARTICLE INFO

Article history:
Received 16 November 2011
Revised 31 January 2012
Accepted 23 February 2012
Available online 1 March 2012

Edited by Judit Ovádi

Keywords:
Chymotrypsin inhibitor 2
Long-range interaction
Network
Nucleation–condensation model
Protein folding
Transition-state

ABSTRACT

The protein folding process has been studied both computationally and experimentally for over 30 years. To date there is no detailed mechanism to explain the formation of long-range interactions between the transition and native states. Long-range interactions are the principle determinants of the tertiary structure. We present a theoretical model which proposes a mechanism for the acquisition of these interactions as they form in a modified version of 'degrees of separation', that we term 'levels of separation'. It is based on the integration of network science and biochemistry.

1. Introduction

Experimental and computational studies have led to the development of several protein folding models. These include most notably, the framework model which depicts folding as a sequential transition with secondary forming first followed by packing into a tertiary structure [1–4] and the nucleation–condensation model [5] which involves both the simultaneous formation of secondary and tertiary interactions in the nucleus. The remaining secondary and tertiary structure then condenses around the nucleus to generate the final folded form. Protein folding also occurs on an energy landscape likened to a folding funnel [6,7]. These models provide significant insight into the way in which proteins fold. However, what appears to be lacking is a detailed mechanism by which native long-range interactions form between the transition and native states. Long-range interactions are defined in this work as contacts between amino acids that are distant in the primary structure ($\geqslant$6) but close in space within the tertiary structure ($\leqslant$5 Å). We propose that this key part of the process occurs through the formation of tertiary interactions radiating from the nucleus in a modified version of the network principle 'degrees of separation' (DOS) [8,9] we term 'levels of separation' (LOS). DOS is defined as the number of connections required to link one amino acid to the rest of the network. Whereas LOS begins not with one amino acid but a small connected subset which then connects to the rest of the network of amino acids in the protein structure.

The application and development of principles from the field of network science has facilitated the study of numerous and disparate systems such as the world-wide web, protein–protein interactions and social networks [10–17]. The study of protein structures and folding has also been significantly advanced by the application of network principles such as the small-world concept which is related to DOS described in this paper [18–29]. In this report we present a theoretical model for the formation of long-range tertiary interactions between the transition and native state. Long-range interactions are the key determinants of the tertiary structure. They are typically defined by distance cutoffs, but are most evident as non-covalent interactions between secondary elements such as β-sheets and α-helices. Because our model protein, chymotrypsin inhibitor 2 (CI2) is very small (64 residues), for clarity we further define long-range interactions as contacts involving residues between β-strands and residues between a β-strand and the α-helix as long as the amino acids involved are six or more residues apart in the primary structure. The calculated network includes interactions mainly involving hydrophobic and van der Waals interactions although there are to a lesser degree hydrogen bonds and ionic interactions present. By evolving the network through the formation of

Abbreviations: CI2, chymotrypsin inhibitor 2; DOS, degrees of separation; LOS, levels of separation; NMR, nuclear magnetic resonance

* Corresponding author. Address: Old Dominion University, Department of Chemistry and Biochemistry, 4541 Hampton Boulevard, Norfolk, VA 23529, USA. Fax: +1 757 683 4628.
E-mail address: lgreene@odu.edu (L.H. Greene).

0014-5793/$36.00
doi:10.1016/j.febslet.2012.02.040

long-range interactions in 'levels of separation' from the experimentally determined nucleus the establishment of the tertiary structure from the transition-state can be for the first time rationalized. This is also particularly important because the study of hydrogen-bonded interactions primarily found in α-helices and β-sheets can be directly monitored kinetically using for example quenched-flow hydrogen-deuterium exchange in conjunction with NMR spectroscopy and folding models developed [30]. However, the study of non-hydrogen-bonded interactions which constitute the majority of long-range interactions is difficult to comprehensively and directly monitor experimentally. Our results show that the protein CI2 can be folded in four LOS through the adaption and application of network principles and macromolecular simulations. This model constitutes a novel and detailed view of an integral aspect of the folding process that has remained largely uncharted.

2. Experimental design to test the hypothesis

2.1. Computational modeling

Given a partially folded structure containing a nucleus we want to create a protein folding landscape using a sequence of modifications of the transition-state structure directed by LOS. The procedure for constructing the structures is as follows: the N_0 network is constructed by taking the nucleus residues and their interactions. Then, we generate an optimal configuration, C_0 , using all interactions associated with the N_0 network. In the next step, we construct N_1 the network by combining the N_0 network and the interacting residues that have a LOS with respect to all the residues in the N_0 network that is equal to one. We use these N_1 network interactions with the C_0 starting configuration to generate an optimal configuration, C_1. The N_2 network is constructed by combining the N_1 network and the interacting residues that have an LOS with respect to all the residues in the N_0 network that is equal to two. We use these N_2 network interactions with the C_1 starting configuration to generate an optimal configuration, C_2. This procedure is continued until we reach the maximum LOS. When we have completed the process, we will have generated a set of networks, called N, and a set of corresponding optimal structural configurations, called C. These sets are defined as

$$N = \{N_0, N_1, N_2, N_3, \ldots, N_j, \ldots, N_{n-1}, N_n\} \quad (1)$$

and

$$C = \{C_0, C_1, C_2, C_3, \ldots, C_j, \ldots, C_{n-1}, C_n\} \quad (2)$$

where the subscripts are associated with the LOS and n is the maximum LOS of all the nucleus residues. We believe that the final structural configuration, C_n, is the native structure.

2.2. Algorithms

Novel algorithms were written in *C* to generate the DOS and LOS in the selected model protein, CI2 based on Eq. 3. The contacts between all atoms were calculated with the program Contact (CCP4) [31]. The contact file was then used as input to calculate all pairwise amino acid long-range contacts with a program we call 'contactToDeglr'. This file was then used as input to calculate the amino acid in each degree of separation with a program we call 'generateDegrees'. All of this information was used as input for a program 'degreesToContacts' to calculate the specific pair-wise interactions for each LOS. Refer to Supplementary Table 1 for the degrees of separation for residues 16, 49 and 57.

2.3. Mathematical modeling: calculating the connectivity at each LOS

A mathematical model has been derived to codify the proposed folding process. This enables us to calculate the number of long-range interactions to be referred to as links, associated with each level in the transition. In a quantitative calculation let m represent the nucleus residues (starting point), let $k_i(m)$ represent the number of residues at the ith LOS from m, and let $l_j(m)$ be the total number of links associated with the jth LOS from m. Then, the total number links, $l_j(m)$, is computed by summing the number of residues starting with the 1st LOS from m up to and including the residues at jth LOS from m. This relationship can be represented mathematically by

$$l_j(m) = \sum_{i=l}^{j} k_i(m) \quad (3)$$

This relationship can also be describe recursively as

$$l_0(m) = k_0$$
$$l_j(m) = k_j(m) + l_{j-1}(m) \quad \text{for } j \geqslant 1 \quad (4)$$

where, k_0 is the total number of residues in the nucleus. For the unsuppressed notation, refer to supplementary material.

2.4. Simulated annealing

The pair-wise interactions for residues in each LOS, was used as restraints in simulations with the program CNS to fold CI2 and initially test the hypothesis [32]. The input file for the simulated annealing process was a completely unfolded linear three-dimensional polypeptide sequence generated in Insight II with Biopolymers (version 2005) (Accelrys, CA). The simulated annealing procedure used 50 000 K for the starting temperature in the high temperature annealing stage and the first slow-cool annealing stage with 1000 or 10 000 steps for each stage. The starting temperature used for the second slow-cooling annealing stage for 3000 steps is 2000 K. 10–20 structures were generated for each simulation. The final minimization stage involved 200 minimization steps and 10 cycles of minimization. Superpositions of the

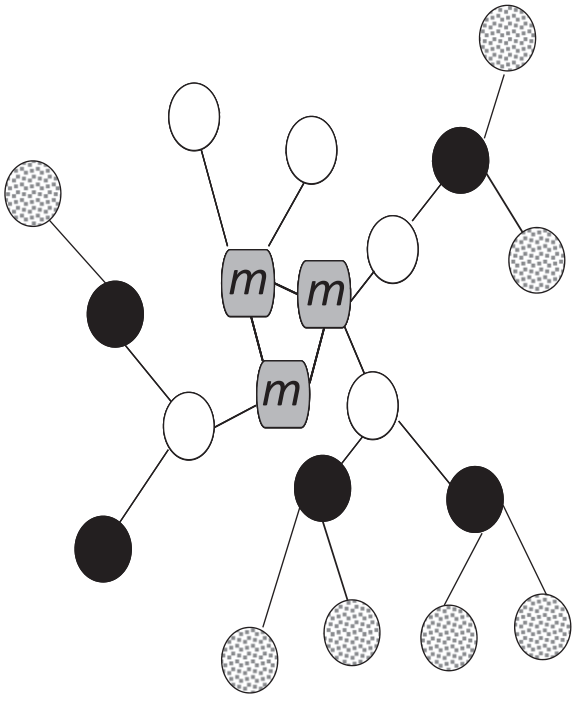

Fig. 1. Simplified network showing levels of separation from a specific node. Here *m*, are the starting nodes. The colors signify LOS. Gray = LOS 0; white = LOS 1; black = LOS 2; dotted = LOS 3. The links between LOS are denoted by black lines.

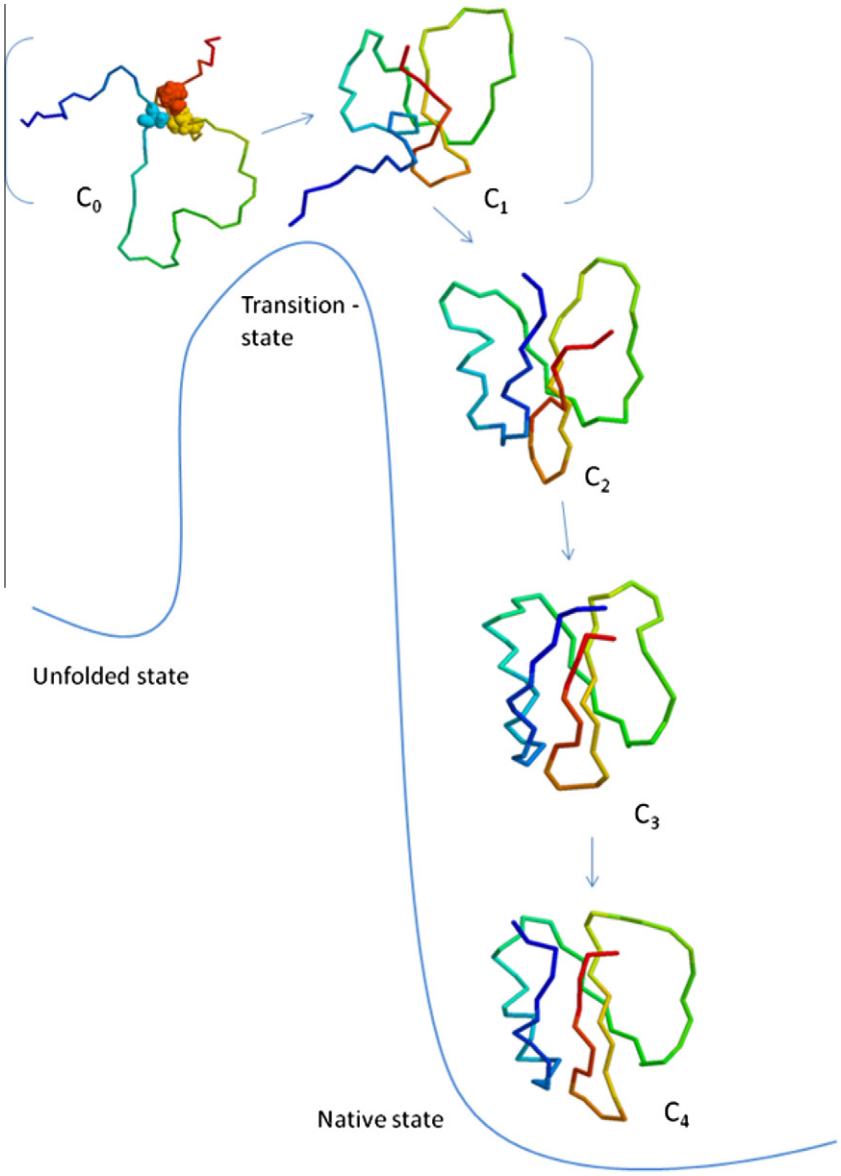

Fig. 2. Folding by levels of separation. Shown are the structures generated using the long-range interactions as restraints in the simulated annealing procedure. The N-terminus is colored blue and the colors transition to the C-terminus in red. The three experimentally determined residues that interact in the nucleus are shown as spheres. The total number of long-range interactions in native CI2 = 124. The total number in each configuration is $C_0 = 3$; $C_1 = 19$; $C_2 = 62$; $C_3 = 100$; $C_4 = 122$. $C_5 = 124$ (data not shown).

native structure of CI2 and the simulated structures generated were conducted with the combinatorial extension program [33].

3. Results and discussion

3.1. Folding by LOS

The folding behavior of CI2 has been studied extensively by Fersht and co-workers using Phi-value analysis [34,35]. This technique enables the indirect elucidation of interactions formed in the transition-state by determining the effects of mutations on stability. In general the folding process of CI2 is cooperative and two-state [5,36]. The nucleation–condensation model however, reveals that CI2 forms a nucleus in the transition state which is comprised of three interacting amino acids (Ala16, Leu49, Ile57) [5,36]. We envision that native interactions form in a cascade radiating from the nucleus enroute to the native state. Through the modification and application of the network principle, DOS [21] into LOS we developed algorithms and a mathematical equation that codifies this process. Computer programs were written which identifies the order of long-range interactions stemming from the nucleus based on this premise and macromolecular simulations were conducted to test this hypothesis and visualize the folding process.

This theoretical model is based on the concept that the initial restriction on conformational space through interactions in the nucleus positions a set of residues in proximity to interact which then further restricts conformational space. This then positions another set of amino acids in proximity to interact with the growing nucleus which continues to restrict conformational space. This process of connectivity proceeds until all native contacts are formed in the native state. Folding by 'levels of separation' is thus cooperative, rapid and reduces the complexity of the search process. Energetically it is downhill because as the number of native interactions increases, the free energy decreases [6,7].

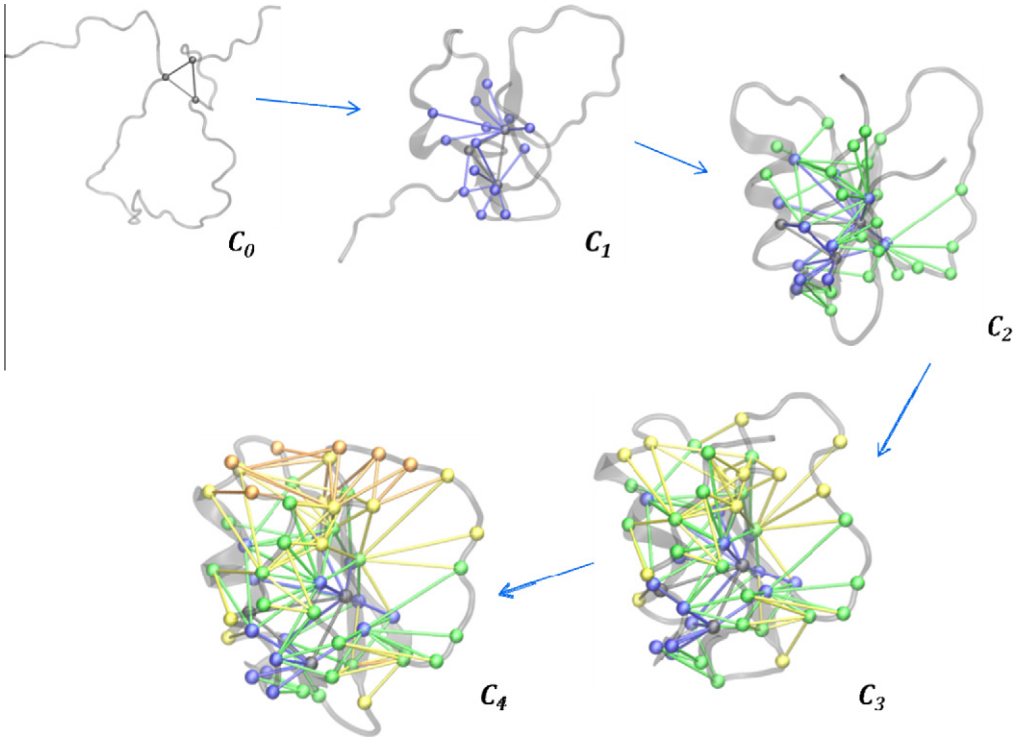

Fig. 3. Visualizing the acquisition of structure by levels of separation. Examples of the long-range interactions used in the simulations are shown in the resultant structures. The addition of these interactions in levels of separations are shown with the following colors: N_0 network (gray); N_1 network (purple); N_2 network (green); N_3 network (yellow); N_4 network (orange).

The network principle, DOS is defined as the measure of how many links it takes to get from one node within a network to another [21]. From here we develop and apply a closely related concept called LOS. This is defined as the number of links to connect an initial group of nodes to the rest of the nodes. In proteins to be assigned an LOS, a residue must be connected to another residue with a finite LOS. In our model, *m*, denotes the residues that are found in the nucleus of the protein (Fig. 1). The residues that are directly connected to *m* have an LOS of 1. The residues that are connected to them (but not to *m*) have an LOS of 2. The residues that are connected to them (but not to *m* or with any residue that has an LOS of 1) have a LOS of 3, and so forth. For each LOS we require the generation of a network which forms the basis for the simulation of a structural configuration.

The network calculated as schematically depicted in Fig. 1 can then be analyzed and used to compute the number of residues and links for each LOS as well as identities of the nodes and relationships in the form of linkage. A quantitative calculation of the number of links and specific residues involved can be found in Figs. 2 and 3 and Supplementary Tables 1 and 2. This is the basis for the algorithm generated to determine the networks.

3.2. Self-organization, the emergence of a giant cluster and the crystallization of a network

The network (N_0) originates with three long-range interactions between Ala16, Leu49, Ile57 in the nucleus and serves as restraints in the calculation of an initial structure, C_0 (Figs. 2 and 3). In the first level of separation sixteen long-range interactions are added to the network (N_0) and a resultant structure C_1 is obtained (Figs. 2 and 3). We envision that these two structures, C_0 and C_1 constitute the transition-state ensemble. Experiments show that the N_0 network is formed in the transition-state but why would the N_1 network be important to the transition-state? C_1 has a gross native-like topology which we propose enables the forming structure to overcome the transition-state barrier (RMSD of 4.2 Å in comparison to the native state). The next network, N_2 is obtained by adding 43 additional long-range interactions to the existing network, N_1 (Figs. 2 and 3). Interestingly, this constitutes 50% of the total number of long-range interactions in the network and the resultant structure is near native. The RMSD between C_2 and the crystal structure is 2.3 Å as calculated with the Combinatorial Extension method [33]. This evokes parallels with several seemingly disparate concepts. For example, Stuart Kauffman proposed previously that evolving networks can become self-organized into a giant component once the number of edges to nodes exceeds 0.5 in his efforts to understand the emergence of biological complexity [37]. The correlation to our protein networks relies first on the conversion of terms whereby a node is an amino acid and an edge is the long-range interaction. We find that when the threshold of the number of links approaches and surpasses 50% the giant cluster to emerge is sufficient to dictate the native-like structure. The remaining two networks and simulated conformations are the result of adding 38 and 22 more long-range interactions to the C_3 and C_4 networks, respectively (Figs. 2 and 3).

These additional interactions appear only to refine the C_2 structure and the RMSD between the simulated structures and the native structure changes from 2.3 Å to 1.0 Å from the C_2 to C_4 structures. We propose that the achievement of C_2 prevents reversal of folding and ensures that the protein continues along the trajectory to the native state. Further, it ensures a smooth landscape. Frustration on the pathway in the context of this hypothesis would then occur when folding does not happen in an orderly LOS. The questions not answered by this work relate to predicting LOS from a nucleus *a priori* without a native state structure and the role of non-native interactions as well as misfolding during the folding process.

Interestingly, if you look at the individual connectivity of each residue it takes from 4 to 7 degrees of separation to connect the

network (Supplementary Fig. 1). There are also a small number of residues, seven, that are not involved in the long-range interaction network. These are residues: 14, 22, 25, 37, 40, 53 and 54.

The concept of percolation theory and the emergence of a giant interconnected cluster within the network after passing a given threshold is also tangentially related to our model [38]. We briefly touch on this with respect to the work of Kauffman discussed earlier [37]. However, in our present system we do not have small unconnected clusters which link into a giant cluster after passing a threshold number which underlies percolation in a network, but instead have a continuous growth of the network stemming from a nucleus and connecting previously unconnected nodes. However, future evolutions of our initial model could involve small clusters of connected nodes outside of the nucleus occurring in larger proteins with simultaneous secondary structure formation. Our concept of LOS also outwardly appears to have parallels with network shells from graph theory. However, our network is not constructed nor analyzed in the same way used to determine *k*-shells. For example, a very common method of analysis of networks is *k*-shell decomposition where, as links are removed *k*-shells are assigned [39,40]. In *k*-shell decomposition a 1-shell is all nodes with one link in the network. When these links are removed a 2-shell are all the nodes with two links. This differs from our approach in that we are not focused on assigning *k*-shells based on number of links but our 'levels' maybe also thought of as 'shells' [40].

4. Conclusion

CI2 can be successfully folded in four LOS starting from the nucleus residues Ala16, Leu49, Ile57 to a native state. This hypothesis provides a conceptual framework to further understand how long-range interactions can form from an initial interconnected nucleus. It also provides further insight into the nucleation–condensation model initially pioneered by Fersht and co-workers. This hypothesis and model is however not predictive and relies on knowledge of the native state structure *a priori*. Perhaps however, this new view will facilitate advancement in the important field of protein structure prediction as well.

Acknowledgements

We are very grateful to Joshua Pothen for valuable assistance with writing the network programs and Jeffrey Tibbitt for expertly drawing Fig. 3. This work was supported by funds from Old Dominion University to L.H.G.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2012.02.040.

References

[1] Santra, M.K., Banerjee, A., Krishnakumar, S.S., Rahaman, O. and Panda, D. (2004) Multiple-probe analysis of folding and unfolding pathways of human serum albumin. Eur. J. Biochem. 271, 1789–1797.
[2] Nolting, B. and Andert, K. (2000) Mechanism of protein folding. Proteins Struct. Funct. Bioinform. 41, 288–298.
[3] Udgaonkar, J.B. and Baldwin, R.L. (1988) NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A. Nature 335, 694–699.
[4] Kim, P.S. and Baldwin, R.L. (1982) Specific intermediates in the folding reactions of small protein and the mechanism of protein folding. Annu. Rev. Biochem. 51, 459–489.
[5] Itzhaki, L.S., Otzen, D.E. and Fersht, A.R. (1995) The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation–condensation mechanism for protein folding. J. Mol. Biol. 254, 260–288.
[6] Socci, N.D., Onuchic, J.N. and Wolynes, P.G. (1998) Protein folding mechanisms and the multidimensional folding funnel. Proteins Struct. Funct. Bioinform. 32, 136–158.
[7] Oliveberg, M. and Wolynes, P.G. (2005) The experimental survey of protein-folding energy landscapes. Q. Rev. Biophys. 38, 245–288.
[8] Watts, D.J. (2003) Six Degrees: The Science of a Connected Age, William Heinmann, London.
[9] Fowler, J.H. and Christakis, N.A. (2008) Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. BMJ, 337.
[10] Schneider, C.M., de Arcangelis, L. and Herrmann, H.J. (2011) Modeling the topology of protein interaction networks. Phys. Rev. E, 84.
[11] Franzosa, E.A. and Xia, Y. (2011) Structural principles within the human-virus protein–protein interaction network. Proc. Natl. Acad. Sci. 108, 10538–10543.
[12] Navlakha, S. and Kingsford, C. (2011) Network archaeology: uncovering ancient networks from present-day Interactions. PLoS Comput. Biol. 7.
[13] Tsai, C.J., Ma, B. and Nussinov, R. (2009) Protein–protein interaction networks: how can a hub protein bind so many different partners? Trends Biochem. Sci. 34, 594–600.
[14] Girvan, M. and Newman, M.E.J. (2002) Community structure in social and biological networks. Proc. Natl. Acad. Sci. 99, 7821–7826.
[15] Newman, M.E.J. and Park, J. (2003) Why social networks are different from other types of networks. Phys. Rev. E, 68.
[16] Bianconi, G. and Barabasi, A.L. (2001) Bose–Einstein condensation in complex networks. Phys. Rev. Lett., 86.
[17] Dezso, Z., Almaas, E., Lukacs, A., Racz, B., Szakadat, I. and Barabasi, A.L. (2006) Dynamics of information access on the web. Phys. Rev. E, 73.
[18] Scala, A., Nunes, Amaral, L.A. and Barthelemy, M. (2001) Small-world networks and the conformation space of a short lattice polymer chain. Europhys. Lett. 55, 594–600.
[19] Vendruscolo, M., Dokholyan, N.V., Paci, E. and Karplus, M. (2002) Small-world view of the amino acids that play a key role in protein folding. Phys. Rev. E, 65.
[20] Dokholyan, N.V., Li, L., Ding, F. and Shakhnovich, E.I. (2002) Topological determinants of protein folding. Proc. Natl. Acad. Sci. USA 99, 8637–8641.
[21] Greene, L.H. and Higman, V.A. (2003) Uncovering network systems within protein structures. J. Mol. Biol. 334, 781–791.
[22] Rao, F. and Caflisch, A. (2004) The protein folding network. J. Mol. Biol. 342, 299–306.
[23] Atilgan, A.R., Akan, P. and Baysal, C. (2004) Small-world communication of residues and significance for protein dynamics. Biophys. J. 86, 85–91.
[24] Bagler, G. and Sinha, S. (2005) Network properties of protein structures. Physica A 346, 27–33.
[25] Brinda, K.V. and Vishveshwara, S. (2005) A network representation of protein structures: implications for protein stability. Biophys. J. 89, 4159–4170.
[26] Higman, V.A. and Greene, L.H. (2006) Elucidation of conserved long-range interaction networks in proteins and their significance in determining protein topology. Phys. A Stat. Mech. Appl. 368, 595–606.
[27] Bode, C., Kovacs, I.A., Szalay, M.S., Palotai, R., Korcsmaros, T. and Csermely, P. (2007) Network analysis of protein dynamics. FEBS Lett. 581, 2776–2782.
[28] Doncheva, N.T., Klein, K., Domingues, F.S. and Albrecht, M. (2011) Analyzing and visualizing residue networks of protein structures. Trends Biochem. Sci. 36, 179–182.
[29] Del Sol, A., Fujihashi, H. and Nusinov, R. (2006) Residues critical for maintaining short paths in network communication mediate signalling in proteins. Mol. Syst. Biol., 2006.
[30] Krishna, M.M.G., Hoang, L., Lin, Y. and Englander, S.W. (2004) Hydrogen exchange methods to study protein folding. Methods 34, 51–64.
[31] Winn, M.D., Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., et al. (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr. Sect. D 67, 235–242.
[32] Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998) Crystallography and NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr. Sect. D, Biol. Crystallogr. 54, 905–921.
[33] Shindyalov, I.N. and Bourne, P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11, 739–747.
[34] Daggett, V. (2006) Protein folding-simulation. Chem. Rev. 106, 1898–1916.
[35] Fersht, A.R. (1995) Optimization of rates of protein folding: the nucleation-condensation mechanism and its implications. Proc. Natl. Acad. Sci. 92, 10869–10873.
[36] Fersht, A.R., Matouschek, A. and Serrano, L. (1992) The folding of an enzyme. I. Theory of protein engineering analysis of stability and pathway of protein folding. J. Mol. Biol. 224, 771–782.
[37] Kauffman, S.A. (1993) The Origins of Life: A New View. Origins or Order: Self-Orgaization and Selection in Evolution, Oxford University Press, NY. pp. 287–342.
[38] Deb, D., Vishveshwara, S. and Vishveshwara, S. (2009) Understanding protein structure from a percolation perspective. Biophys. J. 97, 1787–1794.
[39] Carmi, S., Havlin, S., Kirkpatrick, S., Shavitt, Y. and Shir, E. (2007) A model of internet topology using k-shell decomposition. Proc. Natl. Acad. Sci. USA 104, 11150–11154.
[40] Shao, J., Buldyrev, S.V., Braunstein, L.A., Havlin, S. and Stanley, H.E. (2009) Structure of shells in complex networks. Phys. Rev. E 80, 036105.